# Cooperate without looking: Why we care what people think and not just what they do

# Cooperate without looking: Why we care what people think and not just what they do

Moshe Hoffman[a], Erez Yoeli[a], and Martin A. Nowak[a,b,1]

[a]Program for Evolutionary Dynamics, Department of Mathematics and [b]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138

Edited by Michael S. Gazzaniga, University of California, Santa Barbara, CA, and approved December 24, 2014 (received for review September 30, 2014)

**Evolutionary game theory typically focuses on actions but ignores motives. Here, we introduce a model that takes into account the motive behind the action. A crucial question is why do we trust people more who cooperate without calculating the costs? We propose a game theory model to explain this phenomenon. One player has the option to "look" at the costs of cooperation, and the other player chooses whether to continue the interaction. If it is occasionally very costly for player 1 to cooperate, but defection is harmful for player 2, then cooperation without looking is a subgame perfect equilibrium. This behavior also emerges in population-based processes of learning or evolution. Our theory illuminates a number of key phenomena of human interactions: authentic altruism, why people cooperate intuitively, one-shot cooperation, why friends do not keep track of favors, why we admire principled people, Kant's second formulation of the Categorical Imperative, taboos, and love.**

game theory | evolution | emotion | motive | cooperation

Cooperation occurs when we take on costs to help others. A key mechanism by which cooperation is sustained is reciprocity: Individuals cooperate with those who have cooperated in the past (1–14). However, we care about not only whether others cooperate, but, also, their decision-making process: we place more trust in cooperators who do not strategically weigh the costs and make an effort to collect them before deciding whether to cooperate. For example, we are impressed by colleagues who immediately agree to proofread a paper but view with suspicion those who ask, "how many pages does it have?" Intuitively, those who cooperate without "looking" (CWOL) can be trusted to cooperate even in times when there are large temptations to defect. However, will the added trust from CWOL be worth missing out on those large temptations? Additionally, which conditions make CWOL a winning strategy?

To address these questions, we develop the envelope game (Fig. 1), which is a repeated asymmetric game between two players. In each round, player 1 receives an envelope, which contains the magnitude of the temptation to defect. The temptation is low with probability $p$ and high with probability $1-p$. Player 1 can choose to look inside the envelope and, thus, find out the magnitude of the temptation. Then, player 1 decides to cooperate or defect. Subsequently, player 2 can either continue or end the game. In the former case, there is another round with probability $w$.

If player 1 cooperates, her payoff is $a$, whereas player 2 receives $b$. If player 1 defects, her payoff is either $c_l$ or $c_h$, depending on whether the temptation is low or high, respectively, whereas player 2 receives $d$. We have the following inequalities: $c_h > c_l > a > 0$ and $b > 0 > d$. Moreover, we have $pb + (1-p)d < 0$. Therefore, player 2 prefers not to interact with a player 1 who only cooperates when the temptation is low. Finally, we assume that low temptation is more likely than high temptation: $p > 1/2$.

To understand the essence of the game, we need to consider four strategies for player 1 and three strategies for player 2. The player 1 strategies are (*i*) CWOL, (*ii*) cooperate with looking (CWL), (*iii*) look and cooperate only when the temptation is low, and (*iv*) always defect. The player 2 strategies are (i) end if player 1 looks, (*ii*) end if player 1 defects, and (*iii*) always end. In *SI Appendix*, we also explore a richer strategy set.

The payoff matrix is shown in Table 1. The strategy pair "always defect" and "always end" (ALLD) is always a Nash equilibrium; no player can increase her payoff by deviating unilaterally. However, there are other Nash equilibria. All proofs are in *SI Appendix*.

Of particular interest is the strategy pair where player 1 chooses CWOL and player 2 ends the game if player 1 looks. This strategy pair is a Nash equilibrium if $a/(1-w) \geq c_l p + c_h(1-p)$. This condition has a natural interpretation: player 1's expected temptation from defection is less than the gains from an ongoing cooperative interaction. The expected temptation matters because, if player 1 were to look, player 2 would end the relationship. Thus, player 1 might as well defect, regardless of the temptation. Not looking, in a sense, smooths the temptation to defect; the variability in temptations no longer matters.

Another relevant strategy pair is if player 1 CWL and player 2 ends if player 1 defects. This pair is a Nash equilibrium if $a/(1-w) \geq c_h$. This condition has the following interpretation: to sustain CWL, the long-term gains to player 1 from the ongoing relationship must suffice for player 1 to cooperate, even if player 1 knows the temptation is high in the current period. For CWL, it is the maximal temptation that matters; because player 1 is not penalized for looking, she can look at the temptation and choose to defect only if it is high.

When it is occasionally very costly to cooperate $[c_l p + c_h(1-p) \leq a/(1-w) < c_h]$, CWL is not an equilibrium, but CWOL is. This expression identifies the region where we should be most likely to discover CWOL. In *SI Appendix*, we show that the inequality $pb + (1-p)d < 0$ must hold for CWOL to emerge.

In *SI Appendix*, we also address some concerns. First, after we consider a richer strategy set, there might be other equilibria of the envelope game, in which player 1 sometimes does not look. These equilibria might exist under different conditions from those given above, which may draw into question our statement that looking matters under these conditions. We show that, when we rule out strategies that randomize or depend on the round,

## Significance

**Why do we trust people more when they do good without considering in detail the cost to themselves? People who avoid "looking" at the costs of good acts can be trusted to cooperate in important situations, whereas those who look cannot. We find that evolutionary dynamics can lead to cooperation without looking at costs. Our results illuminate why we attend closely to people's motivations for doing good, as prescribed by deontological ethicists such as Kant, and, also, why we admire principled people, adhere to taboos, and fall in love.**

Author contributions: M.H., E.Y., and M.A.N. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1417904112/-/DCSupplemental.

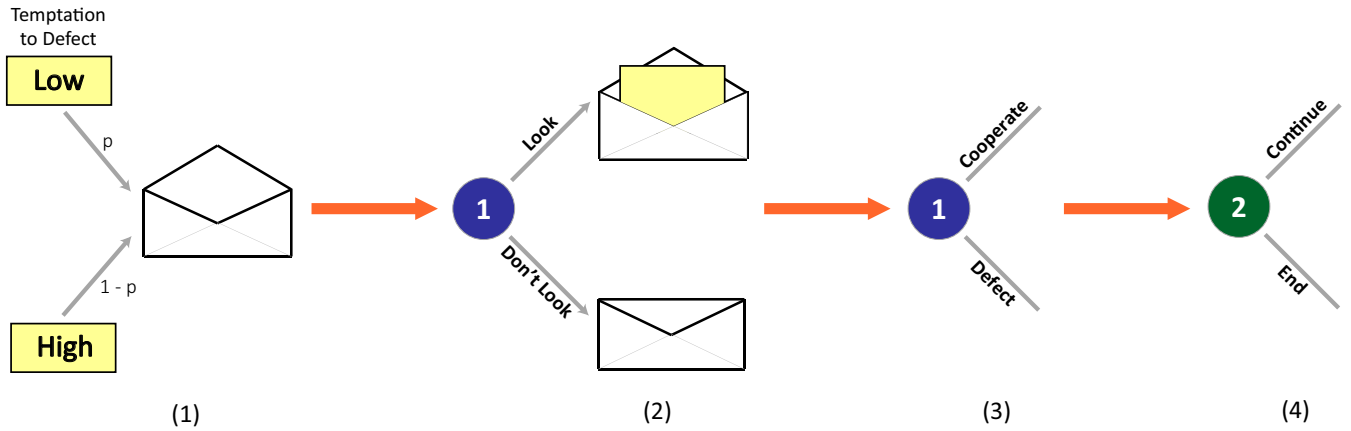


**Fig. 1.** The envelope game. (1) The game begins when the temptation to defect is randomly chosen, as indicated by a notice being placed in the envelope. The temptation to defect is low with probability $p$ and high with probability $1-p$. (2) Player 1 receives the envelope and chooses whether to look (open the envelope). (3) Player 1 decides whether to cooperate or defect. Player 1 can only condition her action on the realized temptation if she has looked. Each time that player 1 cooperates, regardless of whether player 1 looked, both players benefit from the interaction: player 1 gets $a>0$, and player 2 gets $b>0$. Player 1 gains even more if she defects. If the temptation is low, player 1 gets $c_l>a$, and if it is high, player 1 gets $c_h>c_l$. In either case, each time that player 1 defects, player 2 is harmed and gets a negative payoff ($d<0$). Moreover, we assume that the harm is substantial [$d<-bp/(1-p)$], and therefore, player 2 prefers not to interact with a player 1 who only cooperates when the temptation is low. (4) Player 2, having observed both of player 1's choices, decides whether to continue or end. If player 2 continues, there is another round with probability $w$.

ALLD, CWOL, and CWL are the only equilibria of the envelope game, even with this richer strategy set.

Second, because player 2 does not directly benefit by attending to looking, she might not do so. This concern proves moot. The intuition is that, if there is even a small probability that player 1 looks, player 2 is better off attending to looking. We formalize this intuition by showing that this equilibrium, as well as ALLD and CWL, can be made subgame-perfect, which is a solution concept used to rule out these kinds of concerns (15).

In many cases, we do not consciously avoid looking or distrust those who look, but, rather, are guided to do so by a gut sense, an emotion, or an ideology. That is, looking "feels" or "is" wrong. Where do these emotions and ideologies come from? Individuals do not adopt them rationally or even consciously. Therefore, we now consider the case where strategies (such as, it feels wrong to look) are learned or evolved.

We use the replicator dynamic, which is the standard model for evolutionary dynamics (16–18), and also described reinforcement learning and prestige-biased imitation (19). The rate of reproduction is proportional to the payoff that a strategy receives. Because we have two types of players, our simulation studies coevolutionary dynamics in two populations. Players of type 1 can adopt one of four strategies described above. Players of type 2 can adopt one of three strategies described above. Our state space is the product of the simplex $S_4$ and the simplex $S_3$. A point in the simplex $S_4$ describes a strategy mix of type 1 players. A point in the simplex $S_3$ describes a strategy mix of type 2 players.

We randomly seed the strategy frequencies many times and record the frequency of each strategy after the population has stabilized. We observe three possible outcomes that correspond to the Nash equilibria described above (Fig. 2). (*i*) Type 1 players converge to always defect, whereas type 2 players converge to a triangular region close to always end. (*ii*) Type 1 players converge to CWOL, whereas type 2 players converge to a mixture between end if player 1 looks and end if player 1 defects. For stability, this mixture must contain a minimum fraction of end if player 1 looks. (*iii*) Type 1 players converge to a mixture between CWOL and CWL, whereas type 2 players converge to end if player 1 defects. The dynamic stability of those evolutionary outcomes coincides with the criteria for the underlying strategy pairs to be Nash equilibria.

We now apply the model to shed light on some questions directly related to cooperation.

First, psychologists and philosophers have long asked the following question: is helping others "always and exclusively

**Table 1. Payoffs for a restricted set of strategies in the envelope game**

| | Player 2 | | |
|---|---|---|---|
| Player 1 | End if Player 1 looks | End if Player 1 defects | Always end |
| CWOL | $\frac{a}{1-w}, \frac{b}{1-w}$ † | $\frac{a}{1-w}, \frac{b}{1-w}$ | $a,b$ |
| CWL | $a,b$ | $\frac{a}{1-w}, \frac{b}{1-w}$ † | $a,b$ |
| Look and cooperate only when temptation is low | $ap+c_h(1-p), bp+d(1-p)$ | $\frac{ap+c_h(1-p)}{1-pw}, \frac{bp+d(1-p)}{1-pw}$ | $ap+c_h(1-p), bp+d(1-p)$ |
| Always defect | $c_lp+c_h(1-p), d$ | $c_lp+c_h(1-p), d$ | $c_lp+c_h(1-p), d$† |

Player 1's strategies are presented in rows, and player 2's strategies are presented in columns. The payoffs at the intersection of a given row and column are those that the players receive if they play the corresponding strategies. For example, if player 1 looks and cooperates only if the temptation is low and player 2 ends if player 1 defects, then player 1's expected payoff is $[ap+c_h(1-p)]/[1-pw]$, and player 2's expected payoff is $[bp+d(1-p)]/[1-pw]$. Details of calculations leading to payoffs are in *SI Appendix*. Depending on the parameter values, there are up to three Nash equilibria. The pair (ALLD) is always a Nash equilibrium. The pair (CWOL and end if player 1 looks) is a Nash equilibrium if $a/(1-w)>c_lp+c_h(1-p)$. The pair (CWL and end if player 1 defects) is a Nash equilibrium if $a/(1-w)>c_h$. We refer to these strategy pairs as ALLD, CWOL, and CWL, respectively.

†Nash equilibria of the envelope game.

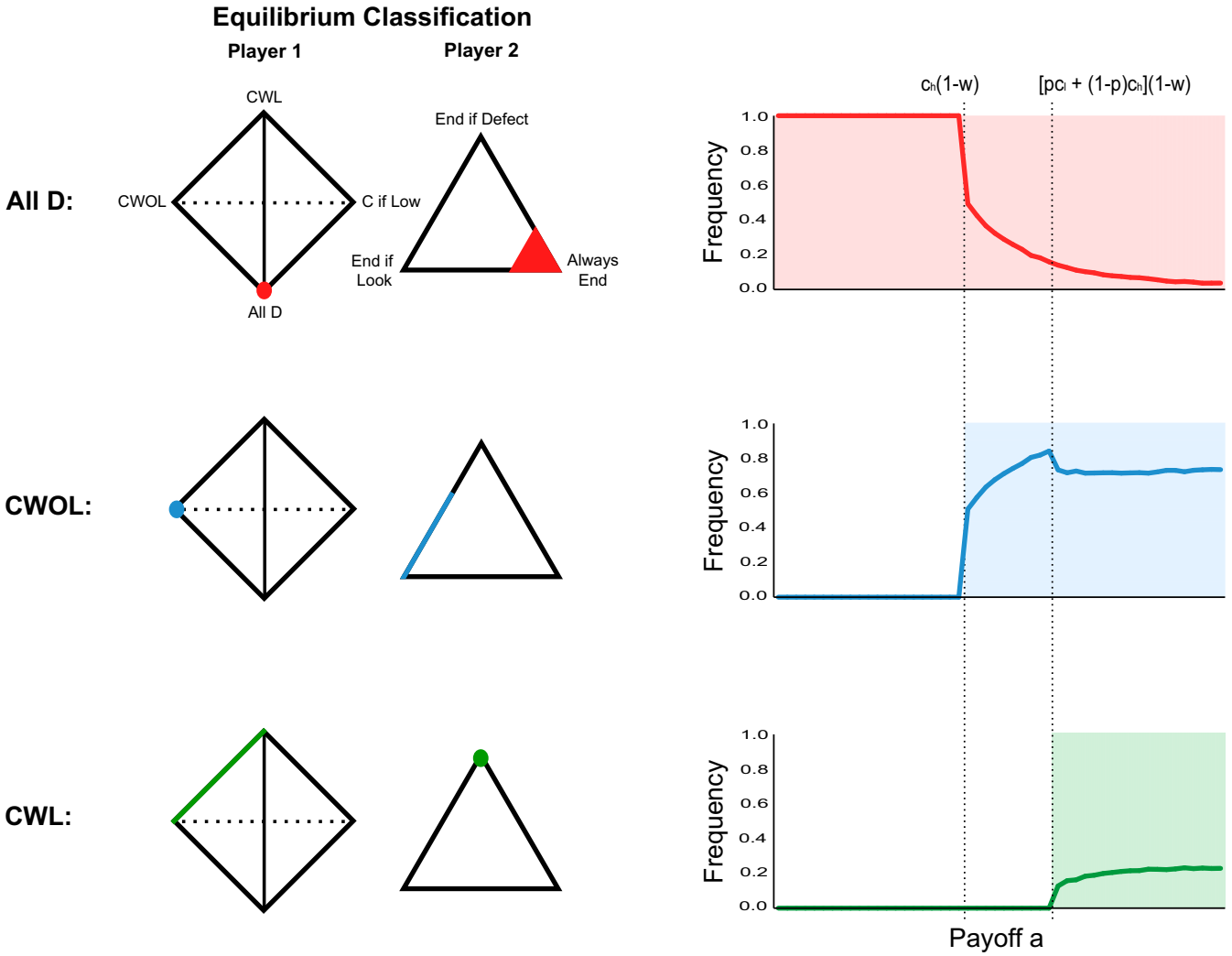


**Fig. 2.** Learning dynamics of the envelope game. We randomly seed the strategy frequencies 10,000 times for 50 values of the payoff value *a* and record the frequency of each strategy after 1,000 generations. We observe three possible outcomes that correspond to the Nash equilibria identified in Table 1. (*i*) Type 1 players converge to always defect, whereas type 2 players converge to a triangular region close to always end. (*ii*) Type 1 players converge to CWOL, whereas type 2 players converge to a mixture between end if player 1 looks and end if player 1 defects. For stability, this mixture must contain a minimum fraction of end if player 1 looks. (*iii*) Type 1 players converge to a mixture between CWOL and CWL, whereas type 2 players converge to end if player 1 defects. We vary the value of *a* along the *x* axis. The *y* axis represents frequencies, and each colored line presents the frequency of each outcome. The parameter region where the corresponding strategy pair is supported as an equilibrium is shaded. Additional details are in *SI Appendix*. All D, all defect; C, cooperate.

motivated by the prospect of some benefit for ourselves, however subtle" (20) [for example, the conscious anticipation of feeling good (21), avoidance of guilt (22–24), reputational benefits, or reciprocity (1–14)]. At the extreme, this question amounts to asking if saintly individuals, such as Gandhi or Mother Teresa, were motivated thus or if they were authentic altruists who did good without anticipating any reward and would be altruistic, even in the absence of such rewards. Our model suggests that authentic altruism is, indeed, possible: by focusing entirely on the benefits to others, authentic altruists are trusted more, and the benefits from this trust outweigh the risk of, for example, dying a martyr's death.

Second, we address another question of why people are intuitive cooperators. That is, when people decide rapidly, they are more likely to cooperate than if they have time to deliberate. Additionally, people who cooperate decide more quickly than those who defect (25–27). Intuitive cooperation underlies extreme acts of heroism, acts which, because they place the hero at great personal risk, are otherwise difficult to explain (28). The Social Heuristics Hypothesis offers one explanation for this phenomenon: we adopt heuristics to avoid incurring cognitive costs associated with deliberation (29–31). In a world with repeated interactions, it is usually worthwhile to cooperate, and therefore, individuals may adopt heuristics, such as always cooperate or always cooperate in nonbusiness settings. These same individuals, when serving as laboratory subjects, may apply these heuristics and cooperate even when it is not worthwhile to do so (25, 32, 33).

Our model offers an additional explanation for intuitive cooperation: intuitive cooperation may serve to reduce responsiveness to costs of cooperating in a particular situation. For this explanation to be sensible, it must be the case that whether a decision is made intuitively or deliberately is detectable. In fact, it is: deliberative decision-making leads to slower reaction time as well as increased pupil size and heart rate (34) and sometimes, blushing or stammering (35). Our model makes two predictions that one would not make if the Social Heuristics Hypothesis, by itself, explained intuitive cooperation: decisions related to cooperation are more likely to be intuitive than other decisions that are similarly usually worthwhile, and intuitive

cooperators are trusted more than reflective cooperators. There is evidence for the latter: in an experiment eliciting moral judgments, subjects who read vignettes about people who returned lost wallets judged those who returned the wallets without hesitation more positively than those who hesitated (36).

Third, we address the question of why people cooperate in one-shot situations (for example, in laboratory experiments, such as the dictator game) (37). Cooperation in these situations is puzzling from the standpoint of models of the evolution of cooperation based on reciprocity, because in that framework, cooperation can only improve one's reputation if actions are observed, and reputations are only valuable if others have an opportunity to reciprocate. Some have suggested that cooperation in these settings results from intuitive cooperation (25, 32), that it can emerge if there is uncertainty over the probability of future cooperation opportunities (38), or that it is a consequence of the evolution of altruism caused by group selection (37). We offer another potential explanation by considering a variation of our model, in which the likelihood of continuation varies and player 1 learns this likelihood when she looks (details in *SI Appendix*). This model suggests that subjects learned or evolved to not consider who is watching so that others can expect them to cooperate, even when no one is. Unlike the other explanations cited, our explanation implies that, when there ends up being an opportunity to interact again, those who cooperate when they thought no one could reciprocate will be rewarded more than those who cooperated knowing someone could reciprocate. Indeed, laboratory subjects cooperate more with those who cooperated with a third party under the presumption that no one would have a chance to reciprocate (39). Moreover, our explanation uniquely predicts that people would feel wrong attending because of the fact that the situation is one shot and that others will judge them harshly if they behave differently when the situation is one shot.

Fourth, we address the question of why we find it unbecoming when close friends keep track of favors or reciprocate favors immediately. In experiments, subjects (*i*) do not cooperate more with friends who have just given them a gift but do cooperate more with strangers who have just given them the same gift (40), (*ii*) take greater care to highlight contributions to strangers than to friends (41), (*iii*) are offended when close friends immediately reciprocate kind acts but not when strangers do (42), and (*iv*) judge friendships as less close when those relationships display immediate reciprocity (43, 44). In fact, relationships have been shown to fall into distinct categories, in part characterized by whether favors are tracked (45). These observations have led researchers to conclude that "the dynamic of friendship does not fit the logic of models of reciprocity and presents a puzzle for evolutionary analysis" (43). However, our model suggests an explanation that is consistent with reciprocity. If close friends CWOL, their decision to cooperate is affected not by a single, recent kind act but, rather, only by the distribution of payoffs from the relationship in the long run. Moreover, when the beneficiary of a good deed immediately reciprocates, then either the beneficiary is looking or the beneficiary thinks that the friend who did the good deed was looking.

Fifth, our model gives insight on a number of interesting phenomena not prima facie related to cooperation.

Why do we like people who are principled and not like those who are strategic? For example, we trust candidates for political office whose policies are the result of their convictions and consistent over time, and we distrust those whose policies are carefully constructed in consultation with their pollsters and who flip-flop in response to public opinion (as caricatured by the infamous 2004 Republican presidential campaign television advertisement showing John Kerry windsurfing and tacking from one direction to the other). Instead of respecting politicians who flexibly respond to public opinion, we view them as sleazy.

Our model offers the following potential explanation. Someone who is strategic considers the costs and benefits to themselves of every decision and will defect when faced with a large temptation, whereas someone who is guided by principles is less sensitive to the costs and benefits to themselves and, thus, less likely to defect. Imagine that our flip-flopping politician was once against gay marriage but supports it now that it is popular. That he only supports it when it is popular indicates that the politician is unlikely to fight for the cause if it later becomes unpopular with constituents or risks losing a big donor. Note that, not only will gay rights activists distrust the flip-flopper but also, women's rights activists will distrust him, even if the flip-flopping politician has always supported women's rights, because the flip-flopper would be likely to end his support for women's issues if it is ever advantageous for him to do so. Of course, we do want our politicians to be strategic about some things. For example, we would prefer that they carefully consider fatalities before invading a foreign country. Our model suggests that we would like politicians—and others more generally—to be strategic about the costs and benefits to us (fatalities) but not the costs and benefits to themselves (likelihood of getting reelected).

Our model also teaches us when we will not be bothered if others are strategic: when defections are either not especially tempting, $a/(1-w) > c_h$, or not especially harmful, $bp + d(1-p) > 0$. Contrast the flip-flopping politician with a business partner who might have the opportunity to cut you out of your latest deal. As long as such a temptation benefits your partner little relative to losing a valuable long-term partnership, your partner would never be tempted, and you need not be bothered if he is strategic.

Next, we discuss why we feel moral disgust by those who use or manipulate others, as famously condemned by Kant in his second formulation of the Categorical Imperative: "Act in such a way that you treat humanity ... never merely as a means to an end, but always at the same time as an end" (46). Consider the well-known example of dwarf-tossing. Many see it as a violation of dwarves' basic dignity to use them as a means for amusement, although dwarves willingly engage in the activity for economic gain. Our aversion to using people may explain many important aspects of our moral intuitions, such as why we judge torture as worse than imprisonment or punishment. Our model suggests that we are repulsed by those who treat others as a means to an end, because they are liable to mistreat their relationship partners when expedient, even if, currently, the relationship is mutually beneficial.

The previous two applications are examples of a more general phenomenon: that we judge the moral worth of an action based on the motivation of the actor as argued by deontological ethicists but contested by consequentialists. The deontological argument is famously invoked by Kant (46):

> Action from duty has its moral worth not in the purpose to be attained by it but in the maxim in accordance with which it is decided upon, and therefore does not depend upon the realization of the object of the action but merely upon the principle of volition in accordance with which the action is done without regard for any object of the faculty of desire.

These applications illustrate that we attend to motives because they provide valuable information on whether the actor can be trusted to treat others well, even when it is not in her interest.

Next, we consider why people dislike considering tradeoffs related to "sacred values" (47). Sacred values are values, such as love, liberty, honor, justice, or life, that people treat "as possessing transcendental significance that precludes comparisons, tradeoffs, or indeed any mingling with secular values" (47). Although there is variation in what societies consider sacred, virtually all societies have a concept of sacredness (47). Sacred values are so strongly imbued in us that we do not find them puzzling prima fascia, but their existence and origin remain poorly understood. What makes us treat some values as sacred,

and what differentiates these values from secular values, like free time or money, that we more readily trade?

Our model provides one possible explanation. People who calculate costs of trading off against sacred values are less trustworthy when it comes to safeguarding these values than people who consider them sacred and would never calculate the costs of trading off against them. Responding with disgust to these taboo tradeoffs may be one way to prevent us from interacting with people who make such tradeoffs and, hence, are less trustworthy and, also, may be a way to signal to others that we ourselves would not consider and, therefore, make such tradeoffs. Consistent with CWOL, it is taboo to consider the tradeoff, even if one ultimately makes the right choice, and the longer the tradeoff is considered for, the harsher the judgment by observers (47). Importantly, those who consider a taboo tradeoff, such as selling their own child, pay a reputational cost, because such considerations indicate that one, in general, does not hold sacred values and cannot be trusted with, for example, care of others' children, the elderly, or shared resources.

If CWOL, indeed, underlies the phenomenon of taboo tradeoffs, then it provides two predictions. First, taboo tradeoffs will prevail precisely in situations where there is large but infrequent temptation to defect and defection is harmful, such as selling a child, betraying a country, or sleeping with someone for a million dollars. It remains to be shown that taboo tradeoffs show these characteristics. Second, it also provides an important policy prescription regarding policies forbidding taboo tradeoffs (for example, the ban on euthanasia): such policies are socially suboptimal, because the benefits of cooperating without looking accrue to the individuals who advocate them, but the costs are borne by society. We note that the above arguments extend to taboos in general and explain why they often have the property that it is not merely a transgression to violate the taboo but to just consider violating it (48).

Finally, our model offers an explanation for emotions, such as love, which is closely related to the explanation first proffered by Frank (49) [precursory insight is given in the work by Schelling (50); also see the works by Hirshleifer (51), Pinker (35), and Winters (52)]. Love has the property that we behave altruistically toward our partners, regardless of what temptations arise (49), as illustrated by the wedding vow "for better or for worse, for richer, for poorer, in sickness and in health." For example, love causes individuals to ignore other potential mates, even if those mates are better than one's current mate, as Shakespeare's Juliet did when her love for Romeo led her to rebuff the advances of the otherwise more suitable Paris.

Why does love have this property? Our model suggests that those who are blinded by love can be trusted to stay with their partners in sickness and health, because they are not looking at the costs of cooperation in these diverse situations. This explanation for love is different from the explanation by Frank (49). Frank (49) argues that those who are blinded by love observably commit to staying with their partners. That is, those who are in love today do not have the option to defect tomorrow. The argument by Frank (49) has been criticized because one could evolve to be in love today and defect tomorrow. Our model requires a different and, we think, more realistic constraint: that it is impossible to look while one appears as though not looking. This assumption is justified by the fact that, at least in some contexts, gathering information about the costs and benefits is inherently observable (for example, through reaction time or the questions that one asks).

Existing evidence is consistent with both models: emotions related to love are observable (53), cannot be faked (54), and are relied on by partners when choosing whether to cooperate (55, 56). There is also reason to believe that love and related emotions would be hard to fake given their autonomic origins and the costs of placing their activation under conscious control (35, 49). It remains to be shown that love, in particular, has these attributes and that we cannot evolve or learn to display love while still attending to costs. Thus, additional research is warranted to differentiate between the model by Frank (49) and CWOL.

Consistent with CWOL, mere discussions of the costs and benefits of a relationship or a breakup (for example, suggesting a prenuptial agreement) damage the relationship. Such discussions indicate that one is looking at the costs of the relationship and cast doubt on one's commitment. CWOL also elucidates that falling in or out of love depends on the distribution of temptations but not their immediate realizations, suggesting that people will fall out of love when there is a permanent change in alternative mating opportunities or relationship costs but not when there is a one-off temptation. For example, one may fall out of love with one's partner after becoming unexpectedly successful. Finally, CWOL clarifies that love comes with a cost—the cost of ignored temptations—and suggests that this cost must be compensated for with commensurate investment in the relationship. Only sometimes is it worthwhile for the recipient of love to compensate a suitor, which explains why people actively avoid the strong affections of those with whom they do not wish to have long-term relationships.

These arguments extend to anger. Anger can be thought of as "punishing without looking." It prevents people from looking at the costs of inflicting harm on others after a transgression, thereby deterring future transgressions.

This paper formalizes a simple intuition first spelled out by Trivers (1):

> One can imagine, for example, compensating for a misdeed without any emotional basis but with a calculating, self-serving motive. Such an individual should be distrusted because the calculating spirit that leads this subtle cheater now to compensate may in the future lead him to cheat when circumstances seem more advantageous (because of unlikelihood of detection, for example, or because the cheated individual is unlikely to survive).

We hope that formalizing this intuition has added valuable insight on otherwise puzzling aspects of human nature.

**ACKNOWLEDGMENTS.** This research was funded, in part, by John Templeton Foundation Grant RFP-12-11 from the Foundational Questions in Evolutionary Biology Fund, National Science Foundation Grant 0905645, and Army Research Office Grant W911NF-11-1-0363.

1. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46(1):35–57.
2. Friedman JW (1971) A non-cooperative equilibrium for supergames. *Rev Econ Stud* 38(1):1–12.
3. Axelrod RM (1984) *The Evolution of Cooperation* (Basic Books, New York).
4. Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54(3):533–554.
5. Fudenberg D, Maskin E (1990) Evolution and cooperation in noisy repeated games. *Am Econ Rev* 80(2):274–279.
6. Binmore KG, Samuelson L (1992) Evolutionary stability in repeated games played by finite automata. *J Econ Theory* 57:278–305.
7. Nowak MA, Sigmund K (1992) Tit for tat in heterogeneous populations. *Nature* 355: 250–253.
8. Nowak M, Sigmund K (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364(6432):56–58.
9. Aumann RJ, Shapley LS (1994) *Long-Term Competition in Game-Theoretic Analysis* (Springer, Berlin).
10. Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393(6685):573–577.
11. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437(7063): 1291–1298.
12. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314(5805): 1560–1563.
13. Ohtsuki H, Iwasa Y (2006) The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol* 239(4):435–444.
14. Sigmund K (2010) *The Calculus of Selfishness* (Princeton Univ Press, Princeton).
15. Osborne MJ (2003) *An Introduction to Game Theory* (Oxford Univ Press, New York).
16. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, United Kingdom).

17. Weibull JW (1997) *Evolutionary Game Theory* (MIT Press, Cambridge, MA).
18. Nowak MA (2006) *Evolutionary Dynamics: Exploring the Equations of Life* (Harvard Univ Press, Cambridge, MA).
19. Fudenberg DA (1998) *The Theory of Learning in Games* (MIT Press, Cambridge, MA), Vol 2.
20. Batson CD (2014) *The Altruism Question: Toward a Social-Psychological Answer* (Psychology Press, New York).
21. Andreoni J (1990) Impure altruism and donations to public goods: A theory of warm-glow giving. *Econ J (London)* 100(401):464–477.
22. Dana J, Cain DM, Dawes RM (2006) What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organ Behav Hum Decis Process* 100:193–201.
23. DellaVigna S, List JA, Malmendier U (2012) Testing for altruism and social pressure in charitable giving. *Q J Econ* 127(1):1–56.
24. Cain DM, Dana J, Newman GE (2014) Giving versus giving in. *Acad Manag Ann* 8(1): 505–533.
25. Rand DG, Greene JD, Nowak MA (2012) Spontaneous giving and calculated greed. *Nature* 489(7416):427–430.
26. Nielsen UH, Tyran JR, Wengström E (2014) Second thoughts on free riding. *Econ Lett* 122(2):136–139.
27. Rubinstein A (2014) A typology of players: Between instinctive and contemplative. Available at www.arielrubinstein.tau.ac.il/papers/Typology.pdf. Accessed January 15, 2015.
28. Rand DG, Epstein ZG (2014) Risking your life without a second thought: Intuitive decision-making and extreme altruism. Available at ssrn.com/abstract=2424036. Accessed January 15, 2015.
29. Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118.
30. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.
31. Kahneman D, Slovic P, Tversky A (1982) *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge Univ Press, Cambridge, United Kingdom).
32. Rand D, et al. (2013) Intuitive cooperation and the social heuristics hypothesis: Evidence from 15 time constraint studies. Available at ssrn.com/abstract=2222683. Accessed January 15, 2015.
33. Rand DG, et al. (2014) Social heuristics shape intuitive cooperation. *Nat Commun* 5(2014):3677.
34. Tursky B, Shapiro D, Crider A, Kahneman D (1969) Pupillary, heart rate, and skin resistance changes during a mental task. *J Exp Psychol* 79(1):164–167.
35. Pinker S (1997) *How the Mind Works* (Norton, New York).
36. Critcher CR, Inbar Y, Pizarro DA (2013) How quick decisions illuminate moral character. *Soc Psychol Personal Sci* 4(3):308–315.
37. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425(6960):785–791.
38. Delton AW, Krasnow MM, Cosmides L, Tooby J (2011) Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proc Natl Acad Sci USA* 108(32):13335–13340.
39. Lin H, Ong D (2011) *Deserving Altruism: Type Preferences in the Laboratory*. Available at ices.gmu.edu/wp-content/uploads/2012/01/Deserving-Altruism-Type-Preferences-in-the-Laboratory-by-Ong-and-Lin.pdf. Accessed January 12, 2015.
40. Boster FJ, Rodriguez JI, Cruz MG, Marshall L (1995) The relative effectiveness of a direct request message and a pregiving message on friends and strangers. *Communic Res* 22(4):475–484.
41. Mills J, Clark MS (1994) Communal and exchange relationships: Controversies and research. *Theoretical Frameworks for Personal Relationships*, eds Erber R, Gilmour R (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ).
42. Shackelford TK, Buss DM (1996) Betrayal in mateships, friendships, and coalitions. *Pers Soc Psychol Bull* 22(11):1151–1164.
43. Silk JB (2003) Cooperation without counting: The puzzle of friendship. *Genetic and Cultural Evolution of Cooperation*, ed Hammerstein P (MIT Press, Cambridge, MA), pp 37–54.
44. Pinker S, Nowak MA, Lee JJ (2008) The logic of indirect speech. *Proc Natl Acad Sci USA* 105(3):833–838.
45. Fiske AP (1992) The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol Rev* 99(4):689–723.
46. Kant I (2002) *Groundwork for the Metaphysics of Morals*, ed Wood AW (Yale Univ Press, New Haven, CT).
47. Tetlock PE (2003) Thinking the unthinkable: Sacred values and taboo cognitions. *Trends Cogn Sci* 7(7):320–324.
48. Fershtman C, Gneezy U, Hoffman M (2011) Taboos and identity: Considering the unthinkable. *Am Econ J Microecon* 3(2):139–164.
49. Frank RH (1988) *Passions Within Reason: The Strategic Role of the Emotions* (WW Norton & Co., New York).
50. Schelling TC (1980) *The Strategy of Conflict* (Harvard Univ Press, Cambridge, MA).
51. Hirshleifer J (1987) *On the Emotions as Guarantors of Threats and Promises* (MIT Press, Cambridge, MA).
52. Winters E (2014) *Feeling Smart: Why Our Emotions Are More Rational Than We Think* (Public Affairs, New York).
53. Ekman P, Sorenson ER, Friesen WV (1969) Pan-cultural elements in facial displays of emotion. *Science* 164(3875):86–88.
54. Ekman P, Davidson RJ, Friesen WV (1990) The Duchenne smile: Emotional expression and brain physiology. II. *J Pers Soc Psychol* 58(2):342–353.
55. Reed LI, Zeglen KN, Schmidt KL (2012) Facial expressions as honest signals of cooperative intent in a one-shot anonymous prisoner's dilemma game. *Evol Hum Behav* 33(3):200–209.
56. Reed LI, DeScioli P, Pinker SA (2014) The commitment function of angry facial expressions. *Psychol Sci* 25(8):1511–1517.